Multiword expressions (MWEs) are diverse and collectively frequent in English. We train a supervised discriminative sequence model on a new annotated corpus to identify heterogenous MWEs in context, giving a lexical semantic segmentation of the sentence. We extend shallow chunking to capture gappy (discontinuous) expressions.

Multiword Expressions

**Definition:** 2 or more space-separated words whose combination is idiomatic in form, function, and/or distribution.

Diverse syntax and semantics:

- Gappy Sequence Tagging

  **Problem:** Identify MWEs as chunks with possible gaps, so as to apply tagging.

  **Solution:** Double the BIO tagset to encode gap status in the state space. Full model: 8 tags

```
They gave me the run around and missing paperwork only to call back to tell me someone else worked there and I would need to come in and put down deposit.
```

Labeled Data

**CMWE**, a text corpus comprehensively annotated with multiword expressions

(Schneider et al., LREC 2014)

- 3,500 manually annotated MWE instances in 3,800 sentences (59k words) of English web reviews
- fully heterogeneous MWEs
- shallow groupings, allowing gaps
- strong (put down) vs. weak (put down–deposit)

Gappy Sequence Tagging

**Problem:** Identify MWEs as chunks with possible gaps, so as to apply tagging.

**Solution:** Double the BIO tagset to encode gap status in the state space. Full model: 8 tags

![Gappy Sequence Tagging](image)

Link-Based Evaluation

Gives partial credit for partial overlap between predicted and gold MWEs. See paper for details.

Experiments

**Preprocessing:** POS tag (retrained TweetNLP tagger on rest of English Web Treebank)

**Model:** First-order structured perceptron tagger (Collins, 2002) with recall-oriented cost to balance recall and precision (Mohit et al., 2012)

**Features:**
- Basic features (summarized below)
- MWE lexicon match
- MWE lexicons extracted from WordNet, SemCor, Prague Czech-English Treebank, SAID, WikiMwe, Wiktionary, and other lists
- Brown clusters from Yelp Academic Dataset

**Baseline:** Match lemmas against lexicons, predict the segmentation with fewest total expressions.

### Basic features adapted from Constant et al. (2012):
- word: current & context, unigrams & bigrams
- POS: current & context, unigrams & bigrams
- capitalization: word shape
- prefixes, suffixes up to 4 characters
- has digit; non-alphanumeric characters
- lemma + context lemma if one is a V and the other is in {N, V, Adj., Adv., Prep., Part.}

Results

Supervised model ≫ non-statistical baseline; lexic matching features help (of {0,2,6,10} lexicons to consult, 6 is best); and:

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Iters</th>
<th>Cost</th>
<th>Params</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>base model</td>
<td>5</td>
<td>1.76k</td>
<td>56.47</td>
<td>58.35</td>
<td>58.35</td>
<td></td>
</tr>
<tr>
<td>+ recall cost</td>
<td>4</td>
<td>150</td>
<td>61.09</td>
<td>59.41</td>
<td>59.41</td>
<td></td>
</tr>
<tr>
<td>+ clusters</td>
<td>3</td>
<td>100</td>
<td>63.98</td>
<td>59.39</td>
<td>59.39</td>
<td></td>
</tr>
<tr>
<td>+ oracle POS</td>
<td>4</td>
<td>2,145</td>
<td>66.19</td>
<td>59.35</td>
<td>62.53</td>
<td></td>
</tr>
</tbody>
</table>

Nathan Schneider · Emily Danchik · Chris Dyer · Noah A. Smith @ Carnegie Mellon University

TACL 2014

Corpus + tagger @ http://www.ark.cs.cmu.edu/LexSem/